5-2006

Analysing Survey Data with Incomplete Responses by Using a Method Based on Empirical Likelihood

Denis H. Y. Leung
Singapore Management University, denisleung@smu.edu.sg

Jing QIN
National Institute of Allergy and Infectious Diseases

DOI: https://doi.org/10.1111/j.1467-9876.2006.00542.x

Published in Journal of the Royal Statistical Society: Series C: Applied Statistics
Volume 55, Issue 3, May 2006, Pages 379-396
https://doi.org/10.1111/j.1467-9876.2006.00542.x

Analysing survey data with incomplete responses by using a method based on empirical likelihood

Denis H. Y. Leung

Singapore Management University, Singapore

and Jing Qin

National Institute of Allergy and Infectious Diseases, Bethesda, USA

Summary. In many surveys, missing response is a common problem. As an example, Zahner, Jacobs, Freeman and Trainor analysed data from a study of child psychopathology in the State of Connecticut, USA. In that study, the response variable, psychopathology, was inferred from questions that were addressed to teachers of the children and was subject to a high level of missingness. However, the missing responses were supplemented by surrogate information that was provided by the parents and/or the primary care providers of the children. In such a situation, it is conceivable that the supplemental information can be used to recover some of the information that has been lost in the cases with missing response. This paper considers a method using empirical likelihood. Empirical likelihood is well known in providing nonparametric inference. But its application has largely been confined to complete-data situations. The method proposed exploits the semiparametric nature of empirical likelihood. The method gives consistent estimates if the cases with non-missing responses form a random sample of the population. In large samples, the method behaves similarly to a regression estimate that is applied to estimating equations. The method is easy to implement with standard statistical packages. In a small sample study, the method was found to give favourable results, when compared with existing methods.

Keywords: Auxiliary information; Empirical likelihood; Missing values; Surrogate; Survey

1. Introduction

Non-response or missing data is a ubiquitous problem in almost all surveys. In some cases, along with the primary response of interest, auxiliary information is also collected. The auxiliary information can be considered to be a cheaper alternative to the primary response and therefore, in the absence of the primary response, can be used as a surrogate. The surrogate, of course, can never be expected to provide the same quality of inference as the primary response. However, it may recover some of the loss of efficiency due to missingness in the primary response.

The motivating example for this paper is a study of psychopathology in urban and rural children in Connecticut (Zahner *et al*., 1993). The study was based on two epidemiological surveys involving a total of 2519 children. The primary goal of the study was to determine whether there were geographical variations in child psychopathology. A rural–urban comparsion is important as it gives researchers an idea of the environmental and social effect on psychological disorders. One of the primary response variables in that study came from the teachers' reports on psychopa-

Address for correspondence: Denis H. Y. Leung, School of Economics and Social Sciences, Singapore Management University, Singapore.
E-mail: denisleung@smu.edu.sg

thology of a child. In that study, 43% of the teachers' reports were missing. However, the investigators also obtained parallel 'parent' reports from one of the parents (or the primary care giver) of the child and, among the 2519 children, only a handful of these parents' reports were missing. In this paper, the data set is reanalysed using a new method. The method proposed uses the parents' reports to supplement information that has been lost through the missing teachers' reports.

In surveys with missing data, the missing data mechanism is important in dictating what methods are suitable for drawing inference. When missingness is independent of any of the variables of interest, then the missing data are said to be missing completely at random (MCAR) (Little and Rubin, 1987). If missingness is dependent only on the observed data, then the missing data are said to be missing at random (MAR). Finally, if missingness is dependent on unobservable data, then the missingness is said to be informative. Under data MCAR, semiparametric methods can be used for making inference, without regard to the missing data mechanism. Furthermore, a 'complete-case' analysis based on only cases with no missing data still provides valid, albeit inefficient, estimates. Under data MAR, the missing data mechanism can still be ignored, if a likelihood-based approach is taken. Therefore, for both data MCAR and data MAR, the missingness mechanism is said to be ignorable. Semiparametric methods can also be used for data MAR situations, but in those cases the missing data mechanism must be taken into account (e.g. Robins *et al*. (1994)). Finally, when missingness is informative, then an unverifiable model for the missing data mechanism must be specified to conduct valid inferences.

The goal of this study is to reanalyse the data from the motivating example by using a semiparametric method. The method proposed requires the data to be MCAR. This choice is inspired by the popularity of the generalized estimating equation model (Liang and Zeger, 1986) in longitudinal analyses, which also requires the assumption of data MCAR. The method proposed here assumes a situation where a survey has been carried out to determine the relationship between a response Y and a set of covariates X. Apart from Y and X, the survey also collected information on a surrogate S of the response. The response Y is missing in a subset of the respondents. In contrast, X and S are measured on all respondents. The relationship between Y and X is assumed to be summarized in a set of estimating equations. When the assumption of data MCAR is satisfied, valid inference can be drawn using the estimating equations based on the complete cases only. However, the incomplete cases also carry information about the relationship between Y and X. The method proposed reweights the estimating equations on the basis of the complete cases by using information from the surrogate. The weights are appropriately determined by the method of empirical likelihood (Owen, 1988). The method proposed is similar to the regression estimate that was suggested by Cochran (1977), except that the regression estimate is based on regressing the mean of Y on that of X, whereas the method that is suggested here is equivalent to regressing estimating equations. In regression settings, the method proposed can be easily implemented with standard statistical packages, once the weights have been obtained. Since the method assumes that the missing reports formed a random sample of the population, a test was carried out to assess the validity of the randomness assumption.

The rest of this paper is organized as follows: Section 2, presentation of the method and its large sample properties; Section 3, the application of the method to the motivating example; Section 4, the results of a simulation study; Section 5, comparison of the method proposed and the existing methods; Section 6, conclusion and discussion.

2. Method proposed and its large sample properties

Assume that Y is the primary response and S is the surrogate, both of which can be continuous or categorical. Furthermore, X is a $1 \times k$ vector representing a set of (continuous or categorical)

covariates. The data are assumed to consist of N observations of which n are complete cases and $N-n$ are cases with missing Y. The data thus comprise $(y_i, s_i, x_i),\ i=1,\ldots,n$ (the validation sample), and $(s_i, x_i),\ i=n+1,\ldots,N$ (the non-validation sample). The goal is to estimate the relationship between Y and X, which can be summarized by a $1\times k$ vector of parameters, β.

The method proposed has its roots in the following problem. Assume that a set of data $(x_1, y_1),\ldots,(x_n, y_n), x_{n+1},\ldots,x_N$ is available, let

$$\bar{y}=\sum_{i=1}^{n} y_i/n,$$

$$\bar{x}=\sum_{i=1}^{n} x_i/n,$$

$$\bar{Y}=\sum_{i=1}^{N} y_i/N,$$

$$\bar{X}=\sum_{i=1}^{N} x_i/N.$$

The interest here is in estimating $\bar{Y}$. Then a naïve method is to use $\bar{y}$ as an estimate of $\bar{Y}$. However, if the auxiliary information can be summarized as $N^{-1}\{w(x_1)+\ldots+w(x_N)\}=0$, for some function w, then Chen and Qin (1993) suggested an estimate that is given by

$$\sum_{i=1}^{n} \tilde{p}_i y_i$$

where the $\tilde{p}_i$s are obtained by maximizing the empirical likelihood

$$\prod_{i=1}^{n} p_i$$

subject to the constraints

$$\sum_{i=1}^{n} p_i=1, \qquad p_i\geqslant 0, \qquad \sum_{i=1}^{n} p_i\, w(x_i)=0.$$

If $w(x_i)=x_i-\bar{X}$, then Chen and Qin's (1993) method is asymptotically equivalent to the well-known regression estimate that is used in survey sampling (Cochran, 1977), i.e.

$$\bar{y}-\sigma_{XY}(\bar{x}-\bar{X}),$$

where σ_{XY} is an estimate of the correlation between X and Y.

Returning to the problem of estimating β, Chen and Qin's (1993) method can be generalized to estimating the relationship between X and Y in the presence of auxiliary information on (S, X). Specifically, it is assumed that

(a) information on β from $(x_1, y_1),\ldots,(x_n, y_n)$ is given by a set of mean 0 estimating equations,

$$\sum_{i=1}^{n} U(y_i, x_i, \beta)=0,$$

(b) auxiliary information is given by a surrogate S and the covariate X in $(x_1, s_1),\ldots,(x_N, s_N)$ and can be summarized by a second set of estimating equations:

$$\sum_{i=1}^{N} R(s_i, x_i, \gamma)=0. \tag{1}$$

In general, R can be a vector; see Sections 4 and 5. If $\hat{\gamma}$ is the solution to equations (1), then β can be estimated by solving the set of weighted estimating equations

$$\sum_{i=1}^{n} \hat{p}_i U(y_i, x_i, \beta) = 0, \tag{2}$$

where $\hat{p}_i$, $i = 1, \ldots, n$, are obtained by maximizing the empirical likelihood

$$\prod_{i=1}^{n} p_i \tag{3}$$

subject to the constraints

$$\sum_{i=1}^{n} p_i = 1, \qquad p_i \geqslant 0, \qquad \sum_{i=1}^{n} p_i R(s_i, x_i, \hat{\gamma}) = 0. \tag{4}$$

Thus, in this approach, the non-validation sample is not used to make direct inference on β. Rather, it is used to obtain the weights $\hat{p}_i$, $i = 1, \ldots, n$, in the set of weighted estimating equations (2).

The optimal values $\hat{p}_i$, $i = 1, \ldots, n$, are obtained by maximizing equations (3) subject to constraints (4). This procedure can be done by introducing Lagrange multipliers, and then profiling the p_is (Owen, 1988) to give

$$\hat{p}_i = \frac{1}{n} \frac{1}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})}, \tag{5}$$

where λ is a $1 \times k$ vector of Lagrange multipliers determined by

$$\sum_{i=1}^{n} \frac{R(s_i, x_i, \hat{\gamma})}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})} = 0. \tag{6}$$

If $\hat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \hat{p}_i U(y_i, x_i, \hat{\beta}) = 0, \tag{7}$$

then for $n \to \infty$ and $n/N \to \rho, 0 < \rho < 1$, $\hat{\beta}$ will have the distributional property

$$(\hat{\beta} - \beta)\sqrt{n} \to \mathrm{MVN}(0, \Sigma) \tag{8}$$

where

$$\Sigma = E\left(\frac{\partial U}{\partial \beta}\right)^{-1} \{V(U) - (1 - \rho)\, E(UR)\, V^{-1}(R)\, E^{\mathrm{T}}(UR)\} \left(E\left(\frac{\partial U}{\partial \beta}\right)^{-1}\right)^{\mathrm{T}}.$$

The large sample property of $\hat{\beta}$ indicates that, asymptotically, using a surrogate always gives more precise estimates. The gain in efficiency is given by the term

$$E(\partial U/\partial \beta)^{-1}(1 - \rho)\, E(UR)\, V^{-1}(R)\, E^{\mathrm{T}}(UR)(E(\partial U/\partial \beta)^{-1})^{\mathrm{T}},$$

which suggests that, everything else being equal, the gain is higher when ρ is smaller. In other words, when the proportion of incomplete cases is large, the gain is high. Also, the expression

$$E(UR)\, V^{-1}(R)\, E^{\mathrm{T}}(UR)$$

represents a 'correlation' between U and R. Therefore, if the information about β that is contained in R is highly correlated with that contained in U, then the gain is higher. This logic in turn means that the gain is higher when the correlation between Y and S is high, since U is a

function of Y, and R is a function of S. The results also suggest that the choice of R will affect the efficiency of the method. Obviously, a poor choice of R leads to a lower correlation between R and U, which means that the information from the surrogate is not extracted efficiently. But, even in that case, it is still better than not using the surrogate at all. The robustness of using the method, in the presence of a weak surrogate, is demonstrated in the simulation results in Sections 4 and 5.

Property (8) also suggests that an α-level confidence region CR_α for β can be constructed by using a Hotelling's T-argument with variance estimated by

$$D_1^{-1} C_{11} D_1^{-1} - (1-\rho) D_1^{-1} C_{12} C_{22}^{-1} C_{12}^{\mathrm{T}} D_1^{-1} \tag{9}$$

where

$$D_1 = n^{-1} \sum_{i=1}^{n} \partial U_i(\hat{\beta}) / \partial \beta,$$

$$C_{11} = n^{-1} \sum_{i=1}^{n} U_i(\hat{\beta})\, U_i^{\mathrm{T}}(\hat{\beta}),$$

$$C_{12} = n^{-1} \sum_{i=1}^{n} U_i(\hat{\beta})\, R_i^{\mathrm{T}}(\hat{\gamma}),$$

$$C_{22} = n^{-1} \sum_{i=1}^{n} R_i(\hat{\gamma})\, R_i^{\mathrm{T}}(\hat{\gamma}).$$

3. A study of rural–urban child psychopathology in Connecticut

Studying the geographical variation in psychopathology is an important research problem. It allows researchers and health care providers to understand the social and environmental influences of psychological disorders. Furthermore, it provides valuable information for policy makers to plan the delivery of the service. Between 1986 and 1989, a study was carried out in Eastern Connecticut, USA, to study this problem in children who were aged between 6 and 11 years (Zahner *et al.*, 1993). The data were collected from two separate surveys: the New Haven Child Survey and the Eastern Connecticut Child Survey. The outcomes in the study were three measures of psychopathology in children: total disturbance and two 'broad-band' measures of emotional ('internalizing') and behavioural ('externalizing') disturbances. In both surveys, the subjects (children) were identified. One of the parents or the primary care provider was then contacted to provide a report on the child. With the consent of the parents and the school-board, the child's teacher was also approached for rating. Information from each parent or primary care provider was gathered by using the child behavioural checklist (CBC) whereas that from the teacher was obtained by using the teachers' report form (TRF) (Achenbach and Edelbrock, 1983, 1986). With over 3500 published studies using the CBC by August 2001 (Achenbach, 2001), the CBC and TRF are arguably the most widely used measures of child psychopathology. Both the CBC and the TRF are continuous scales ranging from 1 to 100, with a higher score on either scale indicating more severe disturbance. CBC and TRF scales were obtained on all three outcomes: total, internalizing and externalizing disturbance. The primary interest was the geographical variation in the three measures of psychopathology that was provided by the TRF. Therefore, the primary outcome in this study was the rating in the teacher's report. The parent's rating served as the surrogate.

Reanalyses on all three measures of psychopathology were carried out. The results are extensive and the trends of the results in the three measures are similar. Therefore, only the results on externalizing are presented. Results on the other two measures can be obtained from the authors on request. The choice of presenting externalizing results is twofold. First, internalizing results have been reported in a few other studies; see, for example, Fitzmaurice *et al.* (1996) and Goldwasser and Fitzmaurice (2001). Second, the use of the parent's rating as a surrogate is more appropriate for the externalizing scale because of the higher correlation between the teacher's and parent's ratings in externalizing (about 0.4) compared with internalizing (about 0.2).

In total, 2519 children were studied, with $N = 2501$ complete parents' (CBC) reports. The missing parents' reports were due to unscorable forms. There were $n = 1433$ children with complete teachers' reports (TRFs) and CBC reports. Most of the missing teachers' reports were because permission was denied by the parents and/or the school-board. The rest were due to non-response. As discussed by Horton and Lipsitz (2001), missingness of this magnitude is quite common in large surveys. The distributions of parents' ratings between the cases with and without missing teachers' reports were similar.

The covariates that were used in the reanalysis were the eight covariates that were used in the original report by Zahner *et al.* (1993). These were the child's sex (CSEX: 1, boy; 0, girl), area (AREA: 1, cities; 0, rural–suburban), social economic status (SES: 1, high; 2, middle; 3, low), single mother (MOMSING: 1, yes; 0, no), mother distress (MOMSTRS: 1, yes; 0, no), family distress (FAMSTRS: 1, yes; 0, no), child with health problems (HLTHPRO: 1, yes; 0, no), child with academic problems (ACADPRO: 1, yes; 0, no). For AREA, 'large cities' and 'small cities' were combined as 'cities', and 'suburban' and 'rural' as 'rural–suburban'. For SES, two binary dummy variables were created, with 'high' as the base-line. Therefore, after recoding, there were a total of nine binary covariates.

The mean externalizing rating in the teachers' reports (TXEXT) was 50.9 (range 39–89) and that in the parents' reports (PXEXT) was 49.3 (range 30–89). A summary of the raw data is given in Table 1. Linear regression was used to model the conditional means of TXEXT and PXEXT. Residuals (which are not shown) from the linear models based on either TXEXT or PXEXT showed no departure from normality.

The method proposed in this paper assumes that the complete cases form a random sample of the population. This assumption is equivalent to data MCAR, in the sense of Little and Rubin (1987). When the assumption of data MCAR fails to hold, the method may give biased parameter estimates. Therefore, the assumption of data MCAR was investigated by comparing the distributions of the variables between the validation and non-validation samples. The results are given in the last three columns of Table 1. The p-values for all variables, except PXEXT, were calculated by using a χ^2-test. For PXEXT, the p-value corresponded to a t-test. It can be observed that, apart from MOMSING and HLTHPRO, the p-values did not reach statistical significance. The 'odds ratio' between the validation and non-validation samples for both MOMSING and HLTHPRO were 1.19. Since such a low magnitude of odds ratio is not indicative of anything of practical significance the analysis therefore seemed to support the assumption of data MCAR.

The application of the method to the data took three steps. In the first step, a standard linear regression of PXEXT on the 10 covariates (nine binary covariates plus an intercept), using all 2501 observations, was carried out. This regression gave the parameter estimate $\hat{\gamma}$ ($\hat{\gamma}$ is a 1×10 vector). This estimate was used in expressions (3) and (4) to obtain the values of $\hat{p}_i, i = 1, \ldots, 1433$. The values of the $\hat{p}_i$s, in the interpretation of weights, should be positive and close to 1 (equal weights). Indeed, for this set of data, $\bar{p} = \Sigma_{i=1}^{n} \hat{p}_i / n = 1.000$ with an interquartile range of 0.070. The minimum value of $\hat{p}_i$ was 0.64 and the maximum value was 1.70. Once the

Table 1. Summary statistics for the data in the psychopathology study

Parameter	*Total (N)*	*%*	*Validation*	*Non-validation*	*p-value*†
AREA					
Cities	1199	47.9	725	577	0.09
Rural–suburban	1302	52.1	708	491	
SES					
High	1240	49.6	728	512	0.35
Middle	949	37.9	528	421	
Low	312	12.5	177	135	
MOMSING					
No	1982	79.2	1161	821	0.02
Yes	519	20.8	272	247	
MOMSTRS					
No	2110	84.4	1199	911	0.3
Yes	391	15.6	234	157	
HLTHPRO					
No	1329	53.1	735	594	0.04
Yes	1172	46.9	698	474	
ACADPRO					
No	1594	63.7	917	677	0.77
Yes	907	36.3	516	391	
CSEX					
Girl	1294	51.7	726	568	0.23
Boy	1207	48.3	707	500	
FAMSTRS					
No	905	36.2	515	390	0.77
Yes	1596	63.8	918	678	
PXEXT					
Mean (standard deviation)			48.97 (10.13)	49.6 (10.49)	0.13

† p-value for the difference between the validation and non-validation samples.

weights had been obtained, they were put into a standard program that can carry out weighted regression to obtain the parameter estimate $\hat{\beta}$.

The results of applying the method proposed (empirical likelihood (EL)) to the data are given in Table 2. Table 2 also gives the maximum likelihood estimator (MLE) of β using only the complete cases (MLE(n)). Using either method, the primary covariate of interest, AREA, is not significant. This result means that, adjusting for other covariates, there is no evidence that there is geographical variation in externalizing disturbance. Among other covariates, only low SES and ACADPRO are significant for both methods. These conclusions are similar to the findings in earlier reports (Zahner *et al*., 1993; Horton and Lipsitz, 2001). There is, however, a difference in conclusion regarding MOMSTRS. The estimate using MLE(n) is 1.659 (standard error SE = 0.679), which is significant at the level $\alpha = 0.05$. Using EL, the parameter estimate is 1.155 (SE = 0.665), which is not significant. For all parameter estimates, the standard error that is given by EL is smaller than the corresponding value given by the complete-case analysis. This finding confirms the results following property (8). However, the reduction in standard

Table 2. Parameter estimates (and standard errors in parentheses) for the adjusted analysis in the psychopathology study

Parameter	*Estimates for the following methods:*	
	MLE (n)	*EL*
CONSTANT	46.725 (0.573)	47.162 (0.567)
AREA	0.452 (0.535)	0.251 (0.523)
LOW SES	1.731 (0.547)	2.054 (0.537)
MIDDLE SES	3.475 (0.883)	4.607 (0.862)
MOMSING	0.625 (0.709)	0.432 (0.698)
MOMSTRS	1.659 (0.679)	1.155 (0.665)
HLTHPRO	−0.136 (0.491)	−0.391 (0.483)
ACADPRO	4.523 (0.525)	3.544 (0.513)
SEX	0.223 (0.492)	0.196 (0.483)
FAMSTRS	1.232 (0.515)	0.966 (0.509)

Table 3. Parameter estimates (and standard errors in parentheses) for the unadjusted analysis in the psychopathology study

Parameter	*Estimates for the following methods:*	
	MLE(n)	*EL*
CONSTANT	50.001 (0.356)	49.840 (0.347)
AREA	1.753 (0.507)	1.754 (0.492)

error from using EL is not substantial, probably because of the moderate correlation between the primary (TXEXT) and surrogate outcome (PXEXT) ($r = 0.37$).

As suggested by Zahner *et al*. (1993), in practice, it is difficult to separate many of the covariates from AREA. Therefore, an unadjusted analysis using only AREA as the covariate is also relevant. The analysis was repeated on the basis of that consideration. The results are given in Table 3. Once again, the estimates by using EL and MLE(n) are similar. There is a slight improvement in precision by EL. The unadjusted analysis now shows a significant elevation of externalizing in the cities. These results are similar to original findings in Zahner *et al*. (1993).

The analytical results in Section 2 show that the effectiveness of EL depends on the correlations between U and R, which in turn depend on the correlation between the primary outcome and the surrogate. In this study, the correlation between the primary outcome (teacher's rating) and the surrogate (parent's rating) is only 0.37. Even though this level of correlation is common in this type of study, according to the meta-analysis of Achenbach *et al.* (1987), nevertheless, the moderate correlation leads to exploration of the possibility of using subsets of the sample that might benefit more from the surrogate data. Findings from previous studies were used to identify such subsets. For example, it has been shown that the correlation tends to be higher in studies involving older children (aged 10 years or older) (Edelbrock *et al*., 1985; Rapee *et al*., 1994; Ollendick *et al*., 2001). It has also been found that lower correlation can be the result of conflicts and stress in the family, particularly if the stress is with the primary care provider

(Jensen *et al.*, 1988; Grills and Ohendick, 2002). Finally, Findling *et al.* (2001) suggested that reports that are done over a long period of time can lead to a lower correlation as it is easy for the child's symptoms to have changed during that time. On the basis of these considerations, subsets of children were identified according to MOMSTRS (= 0), FAMSTRS (= 0), CAGE (child's age, 9–11 years) and TIMELAG (the time between the parent's and teacher's rating less than 15 weeks). The correlation between the surrogate and the primary outcome were found to range between 0.27 and 0.41 in these subsets, which were not significantly higher than the correlation of 0.37 for the entire sample. Hence, there was no evidence that any of these subsets would benefit more from using the surrogate.

4. Simulation results

This section reports the results of a simulation study on the finite sample properties of the method proposed (EL). In the simulation study, the EL method was compared with the (unattainable) MLE by using all observations as if they were all observed (MLE(N)) and the MLE by using the complete cases only (MLE(n)).

Two models, one linear and one non-linear, were considered. The linear model assumed

$$Y = \beta_0 + \beta_1 X + \varepsilon_Y, \qquad \varepsilon_Y, X \sim N(0, 1). \tag{10}$$

Furthermore, S and Y were related by the model

$$S = Y + \varepsilon_S, \qquad \varepsilon_S \sim N(0, \sigma^2). \tag{11}$$

Therefore, S was unbiased for Y but different values of σ would induce different values of the correlation between S and Y.

The non-linear model assumed that Y was a binary 0–1-variable generated by the logistic regression

$$P(Y = 1|X) = 1/\{1 + \exp(-\beta_0 - \beta_1 X)\},$$

where X was a standard normal random variable. S was also binary with

$$P(S = 1) = \eta Y + (1 - \eta)(1 - Y), \qquad 0 \leqslant \eta \leqslant 1. \tag{12}$$

So different values of η would induce different values of the correlation between S and Y.

The following estimating equations were used for the linear model:

$$\begin{aligned} U(Y, X, \beta) &= (Y - \beta_0 - \beta_1 X, (Y - \beta_0 - \beta_1 X)X)^{\mathrm{T}}, \\ R(S, X, \gamma) &= (S - \gamma_0 - \gamma_1 X, (S - \gamma_0 - \gamma_1 X)X)^{\mathrm{T}}. \end{aligned} \tag{13}$$

And, for the non-linear model, the following estimating equations were used:

$$\begin{aligned} U(y, x, \beta) &= (y - p_y, (y - p_y)x)^{\mathrm{T}}, \\ R(s, x, \gamma) &= (s - p_s, (s - p_s)x)^{\mathrm{T}}, \end{aligned}$$

where

$$\begin{aligned} p_y &= 1/\{1 + \exp(-\beta_0 - \beta_1 X)\}, \\ p_s &= 1/\{1 + \exp(-\gamma_0 - \gamma_1 X)\}. \end{aligned}$$

The values of $\beta = (\beta_0, \beta_1) = (0, 1)$ were used in all simulations. Each simulation sample used $N = 300$ cases.

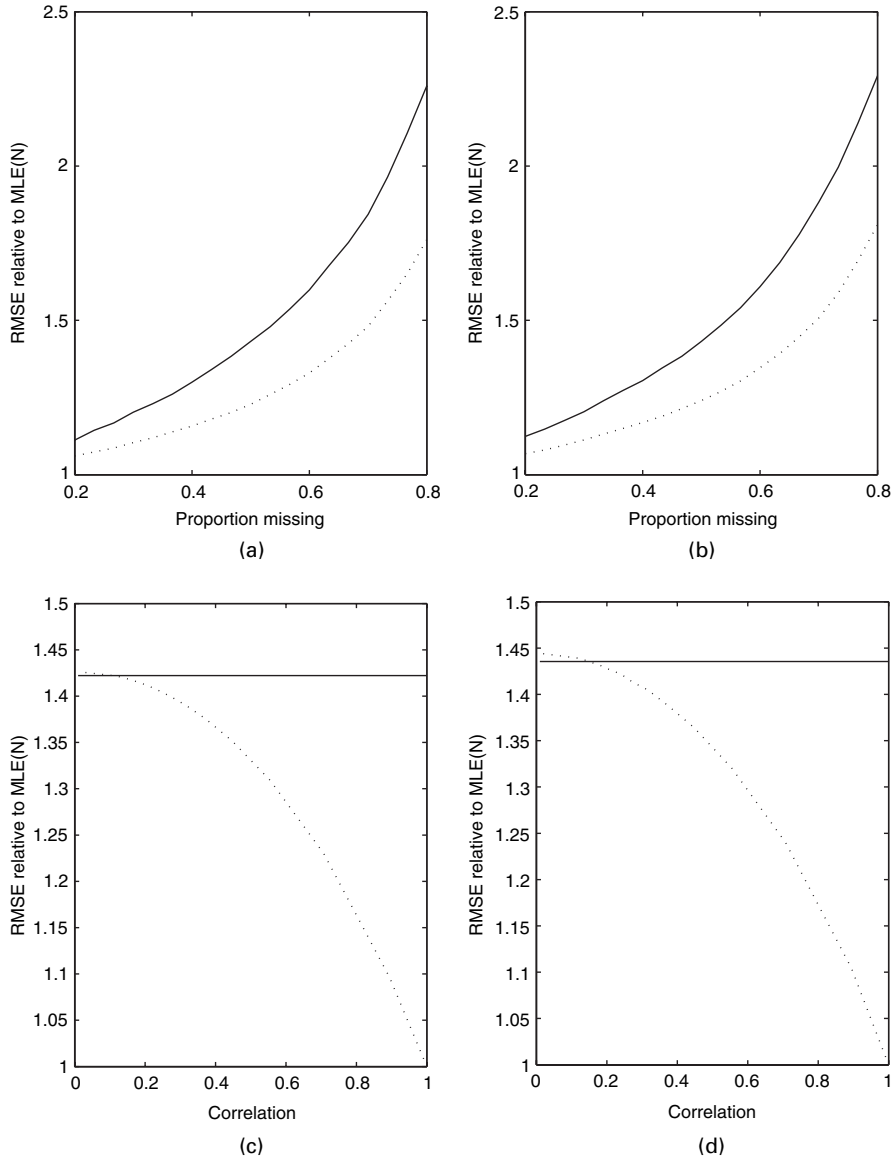

Fig. 1. RMSEs of MLE(n) (———) and EL (· · · · · · ·) relative to MLE(N) as functions of (a), (b) the proportion missing and (c), (d) correlation between Y and S (linear model): (a), (c) relative RMSE for β_0; (b), (d) relative RMSE for β_1

Both the performance of EL and that of MLE(n) depend on the correlation between Y and S and also on the amount of missing data. To study the influence of the proportion of missing data on performance, the relationship between S and Y was fixed at $\sigma = 1$ in model (11) for the linear model and $\eta = 0.7$ in equation (12) for the non-linear model. The relative root-mean-square errors (RMSEs) of MLE(n) and EL to MLE(N) were estimated as the proportion of missing data varied. For each proportion of missingness, 10000 simulation samples were used to estimate the relative RMSEs. To study the influence of the correlation between S and Y on performance, the proportion of missing data was fixed at 50%, so that $n = 150$. The relative

Table 4. Empirical coverage of the 95% EL confidence regions (based on 10 000 simulations)

Correlation	*Results for the following model and values of* $(n, N-n)$*:*			
	Linear model		*Non-linear model*	
	(100,200)	*(150,150)*	*(100,200)*	*(150,150)*
0.8	0.944	0.947	0.951	0.958
0.6	0.937	0.944	0.952	0.956
0.4	0.936	0.941	0.954	0.958
0.2	0.937	0.940	0.954	0.956

RMSEs of MLE(n) and EL to MLE(N) were estimated as the correlation between S and Y varied. The correlation was changed by perturbing the value of σ in model (11) for the linear model and the value of η in equation (12) for the non-linear model. For each correlation value, 10 000 simulation samples were used to estimate the relative RMSEs.

Results of the simulation study for the linear model are summarized in Fig. 1. Figs 1(a) and 1(b) show the performances of MLE(n) and EL as the proportion of missing data ranges from 0.2 to 0.8. The results show that EL is always better than MLE(n) and the relative difference between the two methods increases, in favour of EL, as the proportion of missing data increases. Figs 1(c) and 1(d) show the performances of the two methods as the correlation between S and Y changes from 0 to 1. Since MLE(n) does not use S in making inference, its relative RMSE to that of MLE(N) is independent of the correlation between S and Y. Hence, for MLE(n), the plots of the relative RMSEs for estimating β_0 and β_1 are horizontal lines. The relative RMSEs for EL are quite different. For both β_0 and β_1, the plot drops monotonically as the correlation increases from 0 to 1. Even when the correlation is 0 or close to 0, the relative RMSE for EL is nearly the same as that for MLE(n), for both β_0 and β_1. This finding indicates that, even in the case where S is uninformative about Y, there is little loss in efficiency in using EL. In contrast, when the correlation approaches 1, EL's RMSE approaches that of MLE(N) (the unattainable MLE based on all N cases) and is much smaller than that of MLE(n). The results were similar for the non-linear model case and have been omitted for brevity.

The coverage probabilities of the 95% EL confidence regions for (β_0, β_1) were estimated by using equations (8) and (9) under various combinations of sample sizes and correlation; the results are given in Table 4. The coverage probabilities are very close to the nominal probabilities in all the cases studied. The 90% EL confidence regions were also generated. The results were similar to the 95% cases and are therefore not presented, i.e. the coverage probabilities were close to the nominal level.

5. Comparison with other methods

Several recent works in the literature also aim to supplement missing data with information from surrogate data. All these methods rely on a function of X and S to extract information about β from the non-validation sample. The function $R(S, X, \gamma)$ that is used by EL is an example. Henceforth, in this section, $R(S, X, \gamma)$ is denoted by R_{EL} to distinguish it from the functions that are used by other methods.

In the setting that is considered in this paper, Pepe (1992) suggested maximizing the semiparametric incomplete-data likelihood

$$L(\beta) = \prod_{i=1}^{n} P(y_i|x_i, \beta) \prod_{i=n+1}^{N} P(s_i|x_i, \beta)$$
$$= \prod_{i=1}^{n} P(y_i|x_i, \beta) \prod_{i=n+1}^{N} \int P(y|x_i, \beta)\, P(s_i|x_i, y)\, \mathrm{d}y$$

where $P(s_i|x_i, y)$ is estimated by kernel smoothing. The use of kernel smoothing circumvents the problem of having to specify a parametric model for $P(s_i|x_i, y)$. However, owing to the curse of dimensionality problem that is associated with kernel smoothing, the method is only practical in situations where the dimension of the covariate X is low. For a problem like that considered in this paper, Pepe's (1992) method would be inappropriate. The estimate of β by using Pepe's (1992) approach is found as the solution to the score function

$$\sum_{i=1}^{n} U(y_i, x_i, \beta) + \sum_{i=n+1}^{N} R_{\mathrm{P}}(s_i, x_i),$$

where

$$R_P(s_i, x_i) = \frac{\mathrm{d}}{\mathrm{d}\beta} \int P(y|x_i, \beta)\, P(s_i|x_i, y)\, \mathrm{d}y$$

is the score contribution for the ith observation, $i = n+1, \ldots, N$, calculated over the distribution of Y. This method is sometimes called the mean score imputation method after the fact that the (missing) score contribution for an observation in the non-validation sample is replaced by the mean score R_{P}.

Robins *et al.* (1994) proposed a general class of estimators that in the set-up of this paper is equivalent to finding an estimate of β that is the solution to

$$\sum_{i=1}^{n} \frac{1}{\pi_i} U(y_i, x_i, \beta) + \sum_{i=n+1}^{N} R_{\mathrm{RRZ}}(s_i, x_i),$$

where π_i is the conditional probability that the ith observation belongs to the validation sample, given the covariates. Robins *et al.* (1994) suggested using the data to estimate π_i. For R_{RRZ}, the optimal choice R^* is very complicated and requires knowledge of the unknown probability law generating the data (Chen and Chen, 2000). Chen and Chen (2000) proposed to solve this problem by using a parametric model between S and X. In their method, instead of solving $U(Y, X, \beta)$ for β by using only the validation sample, another estimating equation, say $R_{\mathrm{CC}}(S, X, \eta)$, that summarizes the relationship between the surrogate and the covariates is utilized. If β and η denote the unknown parameters summarizing these relationships, then, under some regularity conditions, the solutions to these equations, $\hat{\beta}$ and $\hat{\eta}$, follow a multivariate normal distribution. The estimate of β that Chen and Chen (2000) proposed is the conditional mean of $\hat{\beta}$ given $\hat{\eta}$ based on the multivariate normality argument. A disadvantage of their method is that a parametric model is required for the relationship between the surrogate and the covariates.

The proposed method (EL) in this paper can be generalized easily to accommodate multiple constraints in expression (4), so in general there can be q constraints, represented as a mean 0 vector, $R(S, X, \gamma) = R_{\mathrm{EL}} = (R_{\mathrm{EL},1}, \ldots, R_{\mathrm{EL},q})^{\mathrm{T}}$. The use of multiple constraints has been discussed in Qin and Lawless (1994) and it gives EL the following advantages.

(a) If one of $R_{\mathrm{EL},1}, \ldots, R_{\mathrm{EL},q}$ is equal to R^*, then EL is fully efficient.

(b) If $R^* = c_0 + \Sigma_{k=1}^{q} c_k h_k$, then using $R_{\mathrm{RRZ}} = (h_1, \ldots, h_k)^{\mathrm{T}}$ in the method of Robins *et al.* (1994) will not give efficient estimates whereas EL based on $R_{\mathrm{EL}} = (h_1, \ldots, h_k)^{\mathrm{T}}$ will still be efficient.

These properties are very attractive. For example, as pointed out by Chen and Chen (2000), it is very difficult to estimate R^*. If $R_{\mathrm{EL},1}, \ldots, R_{\mathrm{EL},q}$ are guesses of R^* then, as long as one of them is correct, EL is fully efficient.

To illustrate how EL compares with earlier methods, a small simulation study was carried out. The simulation study included the methods of Robins *et al.* (1994) and Chen and Chen (2000). The linear model (10), with $\beta = (\beta_0, \beta_1) = (0, 2)$, and $\varepsilon_Y, X \sim N(0, 1)$, was used in the simulations. Four models for S were attempted:

(a) $S = Y$;
(b) $S = Y + \varepsilon_S$, where $\varepsilon_S \sim N(0, 1)$;
(c) $S = 2Y + \varepsilon_S$, where $\varepsilon_S \sim N(0, 1)$;
(d) $S = \sin\{(X-1)/2\} + \varepsilon_S$, where $\varepsilon_S \sim N(0, 1)$.

For Chen and Chen's (2000) method, two choices of R_{CC} were considered:

$$R_{\mathrm{CC}}(S, X, \eta) = (S - \eta_0 - \eta_1 X, (S - \eta_0 - \eta_1 X)X)^{\mathrm{T}},$$

$$R_{\mathrm{CC}}(S, X, \eta) = (S - \eta_0 - \eta_1(0.5X|X|), S - \eta_0 - \eta_1(0.5X|X|)(0.5X|X|))^{\mathrm{T}}.$$

The simulation results corresponding to these two choices of R_{CC} are labelled CC1 and CC2 respectively. Obviously, CC2 is a case of a poor choice of R_{CC}.

For the method of Robins *et al.* (1994), three choices of R_{RRZ} were used. The first assumed that $R_{\mathrm{RRZ}} = 0$. This choice was considered by Rotnitzky and Robins (1995). The second was based on regressing $U(Y, X, \hat{\beta})$ over (X, S, X^2, S^2, XS) using the validation sample, where $\hat{\beta}$ is the parameter estimate that is obtained by solving

$$\sum_{i=1}^{n} \pi_i^{-1} U(y_i, x_i, \beta) = 0.$$

This choice corresponds to the optimal choice R^* under the set-up of the simulations here. The third choice is based on $R_{\mathrm{RZZ}} = R_{\mathrm{CC}}$, as in method CC1. The results corresponding to these choices are labelled RRZ1, RRZ2 and RRZ3 respectively. In addition, when calculating methods RRZ1, RRZ2 and RRZ3, π_i was estimated by using a logistic model: $\mathrm{logit}(\pi_i) = \psi_0 + \psi_1 X$. As Robins *et al.* (1994) pointed out, estimating π_i improves the efficiency.

For EL, two choices of R_{EL} were used: $R_{\mathrm{EL}} = R_{\mathrm{CC}}$ as in model CC1 and $R_{\mathrm{EL}} = R_{\mathrm{RRZ}}$ as in model RRZ2; these are labelled EL1 and EL2 respectively in the simulation results.

For models (a)–(d) for S, 1000 simulation runs were carried out. Each simulation sample used $N = 300$. The results are given in Table 5. Since method RRZ2 used R^*, it is fully efficient and can be used as a bench-mark. However, in practice, method RRZ2 is difficult to achieve. Model (a) corresponds to the case where S is a perfect surrogate. In this case, all methods, except RRZ1 and CC2, are fully efficient. Method RRZ1 is not efficient because it does not use the information from S and method CC2 is not efficient because it uses an inefficient model to summarize the relationship between X and S. Model (b) corresponds to the situation that S is unbiased for Y. In this case, methods EL1, EL2, CC1 and RRZ2 are all efficient. Even though methods CC1, EL1 and RRZ3 use the same function to extract information from the non-validation sample, only methods CC1 and EL1 are efficient. This fact underscores the problem with the method of Robins *et al.* (1994) that was alluded to earlier—if R_{RRZ} and R^* differ by a constant shift or multiple, then the method is no longer efficient. Once again, method RRZ1 is not efficient

Table 5. Bias (and RMSEs in parenthese) of various methods for estimating β_0 and β_1†

Model	*Method*	*Results (× 100) for the following values of n:*			
		n = 150		*n = 100*	
		β_0	β_1	β_0	β_1
(a)	EL1	0.224 (5.96)	0.294 (5.73)	0.209 (5.98)	0.294 (5.83)
	EL2	0.238 (5.94)	0.278 (5.69)	0.238 (5.94)	0.278 (5.67)
	RRZ1	0.265 (8.13)	0.296 (8.14)	0.094 (10.14)	0.441 (9.96)
	RRZ2	0.236 (5.94)	0.293 (5.72)	0.233 (5.95)	0.302 (5.81)
	RRZ3	0.238 (5.94)	0.278 (5.71)	0.237 (5.94)	0.278 (5.71)
	CC1	0.238 (5.94)	0.278 (5.71)	0.238 (5.94)	0.278 (5.71)
	CC2	−1.091 (7.51)	−0.239 (13.52)	0.228 (8.82)	−1.025 (18.36)
(b)	EL1	0.140 (6.98)	−0.235 (7.16)	0.053 (8.21)	−0.379 (8.49)
	EL2	0.145 (6.97)	−0.205 (7.09)	0.061 (8.19)	−0.318 (8.35)
	RRZ1	0.249 (8.28)	−0.304 (8.19)	0.102 (10.26)	−0.430 (10.12)
	RRZ2	0.137 (6.96)	−0.241 (7.15)	0.047 (8.19)	−0.403 (8.47)
	RRZ3	0.009 (8.05)	−0.183 (8.27)	−0.008 (9.70)	−0.352 (10.15)
	CC1	0.144 (6.97)	−0.214 (7.09)	0.061 (8.23)	−0.342 (8.34)
	CC2	0.267 (7.58)	0.322 (13.05)	0.299 (9.24)	0.753 (18.12)
(c)	EL1	−0.354 (6.18)	0.173 (6.25)	−0.199 (6.84)	0.051 (6.89)
	EL2	−0.372 (6.16)	0.186 (6.18)	−0.232 (6.75)	0.070 (6.73)
	RRZ1	−0.368 (8.15)	−0.090 (7.83)	0.041 (10.13)	−0.342 (9.92)
	RRZ2	−0.361 (6.18)	0.182 (6.24)	−0.202 (6.80)	0.079 (6.87)
	RRZ3	−0.389 (10.04)	0.553 (10.23)	−0.664 (13.07)	0.728 (13.06)
	CC1	−0.376 (6.16)	0.179 (6.22)	−0.236 (6.77)	0.086 (6.78)
	CC2	−0.148 (7.41)	−0.440 (13.62)	0.042 (8.88)	−0.174 (18.98)
(d)	EL1	−0.289 (12.18)	−0.483 (8.50)	−0.193 (14.93)	−0.475 (10.28)
	EL2	−0.287 (11.94)	−0.521 (8.37)	−0.254 (14.59)	−0.564 (10.11)
	RRZ1	−0.249 (11.92)	−0.489 (8.32)	−0.241 (14.56)	−0.476 (10.07)
	RRZ2	−0.285 (12.17)	−0.480 (8.50)	−0.235 (14.91)	−0.497 (10.31)
	RRZ3	−0.022 (14.35)	−0.349 (10.21)	−0.172 (18.49)	−0.425 (13.27)
	CC1	−0.563 (15.57)	−0.750 (12.80)	−0.710 (21.03)	−0.921 (17.45)
	CC2	−0.543 (26.16)	−0.291 (25.97)	−1.887 (34.67)	−1.163 (33.69)

†Based on 1000 simulations.

because it omits the information from S in the non-validation sample. For model (c), methods EL1, EL2, CC1 and RRZ2 are once again efficient. Method RRZ3's relative efficiency to the efficient estimates is even lower; methods RRZ1 and CC2 are inefficient for the same reason as in the previous models. Model (d) deals with the case of a weak surrogate. In this case, the performances of methods EL1, EL2, RRZ1 and RRZ2 are similar; method CC1's relative efficiency is much lower.

6. Discussion

In regression settings, the method that is proposed in this paper can be implemented quite easily with any standard statistical package that is capable of carrying out regression analyses. The only adjustment that is needed for this implementation is the input of the weights $\hat{p}_i$. Once the weights have been obtained, standard statistical packages that can handle weighted regressions can be used to obtain the final estimates. As such, the method can be used even if there are a large number of covariates, such as in the psychopathology study. As demonstrated in Section 4, the method is applicable to both discrete and continuous responses.

The simulation results in Section 5 show that, when compared with other methods that have appeared in the literature, the method proposed is very competitive.

The method can be easily extended to handle situations where surrogates for both the response and the covariates are available. In those cases, the function R would be in terms of a surrogate for the response and surrogates for the covariates.

The set-up in this paper assumes that the non-validation sample forms a random sample from the population. This restriction is equivalent to the assumption that data are MCAR. However, the method that is proposed here can be easily modified to accommodate situations where data are MAR. Such situations occur, for example, in two-phase sampling, where in the first phase a surrogate that is measured in the first phase is used to select a subset for the second phase where the primary response is measured and, given the surrogate, the primary response is independent of whether an observation is selected for the second phase. In such situations, the method proposed still provides valid inference as long as a correct model for the probability of selection for the second phase is available and used as one of the constraints in expression (4).

Several earlier works have considered the data that were studied in this paper. Horton and Fitzmaurice (2002) and Ibrahim *et al*. (2001) used a binary response for both the teachers' and the parents' reports (non-clinical *versus* clinical). Both assumed that missingness was non-ignorable. In Ibrahim *et al*. (2001), the missing primary response was 'filled in' by using auxiliary data. The key requirement in the method of Ibrahim *et al*. (2001) is the availability of auxiliary data that are highly correlated with the unobserved data, so that, given the auxiliary data, the missing data mechanism mimics data MAR. Horton and Fitzmaurice (2002) considered the teacher and parent ratings as bivariate responses and proposed the use of a mixture model for drawing inference. The idea of the mixture model is to decompose the joint distribution of the data and the missingness patterns into a conditional distribution of the data given the patterns of missingness and a marginal distribution of the patterns of missingness. Therefore, in a mixture model, there is a natural allowance for different patterns of missingness. Horton and Fitzmaurice (2002) used this fact to accommodate the possibility that some of the missing data can be ignorable whereas others are non-ignorable. Fitzmaurice *et al*. (1996) and Goldwasser and Fitzmaurice (2001) modelled the teachers' and parents' reports as multivariate responses and used likelihood-based regression models. Finally, Horton and Laird (2001) studied the problem of augmenting missingness in the covariates by auxiliary information.

Acknowledgements

We thank Dr G. Zahner for providing the data set in this paper. The data were collected under contract to the Connecticut Department of Children and Youth. We are grateful to Dr N. Horton for his generous help in obtaining the data and for his valuable comments and suggestions. We also thank Dr D. J. Dekle for carefully reading through the paper. The comments from the referees, the Associate Editor and the Joint Editor have been most valuable. This research was partially supported by grants from the Wharton–Singapore Management University Research Center at Singapore Management University.

Appendix A: Derivations of the large sample properties (8) of $\hat{\beta}$

The primary and the surrogate response are denoted by Y and S respectively. Given the k-dimensional covariate $X = (X_0, X_1, \ldots, X_{k-1})^{\mathrm{T}}$, the distribution of Y becomes parameterized by the k parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})^{\mathrm{T}}$. Allowing $n \to \infty$ and $n/N \to \rho$, and

$$R^{\mathrm{T}}(s_i, x_i, \beta) = (R_0(s_i, x_i, \beta), R_1(s_i, x_i, \beta), \ldots, R_{k-1}(s_i, x_i, \beta)),$$
$$U^{\mathrm{T}}(y_i, x_i, \beta) = (U_0(y_i, x_i, \beta), U_1(y_i, x_i, \beta), \ldots, U_{k-1}(y_i, x_i, \beta))$$

are zero-mean estimating equations. From equations (5) and (6), $\hat{p}_i$, $i = 1, \ldots, n$, are determined by

$$\hat{p}_i = \frac{1}{n} \frac{1}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})}, \tag{14}$$

where $\lambda = (\lambda_0, \ldots, \lambda_{p-1})^{\mathrm{T}}$ are Lagrange multipliers determined by

$$\sum_{i=1}^{n} \frac{R(s_i, x_i, \hat{\gamma})}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})} = 0. \tag{15}$$

It follows from equation (15) that

$$\sum_{i=1}^{n} R(s_i, x_i, \hat{\gamma}) \{1 - \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})\} + o_p(n^{-1/2}) = 0.$$

On omitting terms that are higher than $||\lambda||$, this implies that

$$\begin{aligned}
\lambda &= \left\{ \frac{1}{n} \sum_{i=1}^{n} R(s_i, x_i, \hat{\gamma})\, R(s_i, x_i, \hat{\gamma})^{\mathrm{T}} \right\}^{-1} \frac{1}{n} \sum_{i=1}^{n} R(s_i, x_i, \hat{\gamma}) + o_p(n^{-1/2}) \\
&= V^{-1}(R) \frac{1}{n} \sum_{i=1}^{n} R(s_i, x_i, \hat{\gamma}) + o_p(n^{-1/2}) \\
&= V^{-1}(R) \frac{1}{n} \sum_{i=1}^{n} \left\{ R(s_i, x_i, \hat{\gamma}) - \frac{1}{N} \sum_{i=1}^{N} R(s_i, x_i, \hat{\gamma}) \right\} + o_p(n^{-1/2}) \\
&= V^{-1}(R) \frac{1}{n} \left\{ (1-\rho) \sum_{i=1}^{n} R(s_i, x_i, \hat{\gamma}) - \rho \sum_{i=n+1}^{N} R(s_i, x_i, \hat{\gamma}) \right\} + o_p(n^{-1/2}) \\
&= V^{-1}(R) \left[\frac{1-\rho}{n} \sum_{i=1}^{n} \left\{ R(s_i, x_i, \beta) + (\hat{\gamma} - \gamma) \frac{\partial R(s_i, x_i, \beta)}{\partial \beta} \right\} \right. \\
&\quad \left. - \frac{\rho}{n} \sum_{i=n+1}^{N} \left\{ R(s_i, x_i, \beta) + (\hat{\gamma} - \gamma) \frac{\partial R(s_i, x_i, \beta)}{\partial \beta} \right\} \right] + o_p(n^{-1/2}) \\
&= V^{-1}(R) \left\{ \frac{1-\rho}{n} \sum_{i=1}^{n} R(s_i, x_i, \beta) - \frac{\rho}{n} \sum_{i=n+1}^{N} R(s_i, x_i, \beta) \right\} + o_p(n^{-1/2}).
\end{aligned} \tag{16}$$

Also, from equation (7), $\hat{\beta}$ is a solution of

$$\sum_{i=1}^{n} \frac{U(y_i, x_i, \hat{\beta})}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})} = 0. \tag{17}$$

Expanding the numerator of equation (17) in a neighbourhood of β, and rearranging,

$$\begin{aligned}
\hat{\beta} - \beta &= -\left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial U(y_i, x_i, \beta)/\partial \beta}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\beta})} \right\}^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{U(y_i, x_i, \beta)}{1 + \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})} + o_p(n^{-1/2}) \\
&= -E\left(\frac{\partial U}{\partial \beta} \right)^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} U(y_i, x_i, \beta) \{1 - \lambda^{\mathrm{T}} R(s_i, x_i, \hat{\gamma})\} \right] + o_p(n^{-1/2}) \\
&= -E\left(\frac{\partial U}{\partial \beta} \right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} U(y_i, x_i, \beta) - \lambda^{\mathrm{T}} E(UR) \right\} + o_p(n^{-1/2}) \\
&= -E\left(\frac{\partial U}{\partial \beta} \right)^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} U(y_i, x_i, \beta) - E(UR) V^{-1}(R) \frac{1-\rho}{n} \sum_{i=1}^{n} R(s_i, x_i, \beta) \right. \\
&\quad \left. + E(UR)\, V^{-1}(R) \frac{\rho}{n} \sum_{i=n+1}^{N} R(s_i, x_i, \beta) \right\} + o_p(n^{-1/2})
\end{aligned}$$

$$= -E\left(\frac{\partial U}{\partial \beta}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\{U(y_i, x_i, \beta) - E(UR)\,V^{-1}(R)(1-\rho)\,R(s_i, x_i, \beta)\}\right.$$
$$\left. + E(UR)\,V^{-1}(R)(1-\rho)\frac{1}{N-n}\sum_{i=n+1}^{N} R(s_i, x_i, \beta)\right] + o_p(n^{-1/2})$$
$$= -E\left(\frac{\partial U}{\partial \beta}\right)^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}[U(y_i, x_i, \beta)\,E(UR)\,V^{-1}(R)(1-\rho)\{R(s_i, x_i, \beta) - E(R)\}]\right.$$
$$\left. + E(UR)\,V^{-1}(R)(1-\rho)\frac{1}{N-n}\sum_{i=n+1}^{N}\{R(s_i, x_i, \beta) - E(R)\}\right] + o_p(n^{-1/2}).$$

It can easily be shown that, asymptotically,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[U(y_i, x_i, \beta) - E(UR)\,V^{-1}(R)(1-\rho)\{R(s_i, x_i, \beta) - E(R)\}] \to \mathrm{MVN}(0, \Sigma_1),$$

where

$$\Sigma_1 = V(U) - (1-\rho)(1+\rho)\,E(UR)\,V^{-1}(R)\,E^{\mathrm{T}}(UR).$$

Also

$$E(UR)\,V^{-1}(R)(1-\rho)\frac{1}{N-n}\sum_{i=n+1}^{N}\{R(s_i, x_i, \beta) - E(R)\} \to \mathrm{MVN}(0, \Sigma_2),$$

where

$$\Sigma_2 = \rho(1-\rho)\,E(UR)\,V^{-1}(R)\,E^{\mathrm{T}}(UR).$$

Therefore,

$$(\hat{\beta} - \beta)\sqrt{n} \to \mathrm{MVN}(0, \Sigma),$$

where

$$\Sigma = E\left(\frac{\partial U}{\partial \beta}\right)^{-1}\{V(U) - (1-\rho)\,E(UR)\,V^{-1}(R)\,E^{\mathrm{T}}(UR)\}\left(E\left(\frac{\partial U}{\partial \beta}\right)^{-1}\right)^{\mathrm{T}}.$$

References

Achenbach, T. M. (2001) Bibliography of published studies using ASEBA instruments. University of Vermont, Burlington. (Available from `www.aseba.org`.)

Achenbach, T. M. and Edelbrock, C. S. (1983) *Manual for the Child Behavior Checklist and the Revised Child Behavior Profile*. Burlington: University of Vermont Department of Psychiatry.

Achenbach, T. M. and Edelbrock, C. S. (1986) *Manual for the Teacher Report Form and Teacher Version of the Child Profile*. Burlington: University of Vermont Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H. and Howell, C. T. (1987) Child/adolescent behavioral and emotional problems implications of cross-informant correlations for situational specificity. *Psychol. Bull.*, **101**, 213–232.

Chen, J. and Qin, J. (1993) Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, **80**, 107–116.

Chen, Y.-H. and Chen, H. (2000) A unified approach to regression analysis under double-sampling designs. *J. R. Statist. Soc.* B, **62**, 449–460.

Cochran, W. G. (1977) *Sampling Techniques*. New York: Wiley.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R. and Conover, N. C. (1985) Age differences in the reliability of the psychiatric interview of the child. *Child Devlpmnt*, **56**, 265–275.

Findling, R. L., Gracious, B. L., McNamara, N. K., Youngstrom, E. A., Demeter, C. and Calabrese, J. R. (2001) Rapid continuous cycling and psychiatric co-morbidity in pediatric biploar I disorder. *Bipol. Disordr.*, **3**, 202–210.

Fitzmaurice, G. M., Laird, N. M. and Zahner, G. (1996) Multivariate logistic models for incomplete binary responses. *J. Am. Statist. Ass.*, **91**, 99–108.

Goldwasser, M. and Fitzmaurice, G. M. (2001) Multivariate linear regression of childhood psychopathology using multiple informant data. *Int. J. Meth. Psychol. Res.*, **10**, 1–10.

Grills, A. E. and Ohendick, T. H. (2002) Issues in parent-child agreement: the case of structure diagnostic interviews. *Clin. Chld Fam. Psychol. Rev.*, **5**, 57–83.
Horton, N. J. and Fitzmaurice, G. M. (2002) Maximum likelihood estimation of bivariate logistic models for incomplete responses with indicators of ignorable and non-ignorable missingness. *Appl. Statist.*, **51**, 281–295.
Horton, N. J. and Laird, N. M. (2001) Maximum likelihood analysis of logistic regression models with incomplete covariate data and auxiliary information. *Biometrics*, **57**, 34–42.
Horton, N. J. and Lipsitz, S. (2001) Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am. Statistn*, **55**, 244–254.
Ibrahim, J. G., Lipsitz, S. R. and Horton, N. (2001) Using auxiliary data for parameter estimation with non-ignorably missing outcomes. *Appl. Statist.*, **50**, 361–373.
Jensen, P. S. Traylor, J., Xenakis, S. N. and Davis, H. (1988) Child psychopathology rating scales and interrater agreement, I: parents' gender and psychiatric symptoms. *J. Am. Acad. Chld Adolesc. Psychiatr.*, **27**, 442–450.
Liang, K. Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.
Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: Wiley.
Ollendick, T. H., Grills, A. E. and King, N. J. (2001) Applying developmental theory to the assessment and treatment of childhood disorders: does it make a difference? *Clin. Psychol. Psychther.*, **8**, 304–314.
Owen, A. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.
Pepe, M. S. (1992) Inference using surrogate outcome data and a validation sample. *Biometrika*, **79**, 355–365.
Qin, J. and Lawless, J. (1994) Empirical likelihood and general estimating functions. *Ann. Statist.*, **22**, 300–325.
Rapee, R. M., Battett, P. M., Dadds, M. R. and Evans, L. (1994) Reliability of the DSM-III-R childhood anxiety disorders using structured interview: interrater and parent-child agreement. *J. Am. Acad. Chld Adolesc. Psychiatr.*, **33**, 984–992.
Robins, J. M., Rotnizky, A. and Zhao, L. P. (1994) Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.*, **89**, 846–866.
Rotnitzky, A. and Robins, J. M. (1995) Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, **82**, 805–820.
Zahner, G., Jacobs, J., Freeman, D. H. and Trainor, K. F. (1993) Rural-urban children psychopathology in a Northeastern US State: 1986-1989. *J. Am. Acad. Chld Adolesc. Psychiatr.*, **32**, 378–387.